3K

УДК 004.934

Ю.Г. Кривонос, А.С. Загваздин, Ю.В. Крак
Институт кибернетики НАН Украины им. В.М. Глушкова, г. Киев
alex.zagvazdin@gmail.com, krak@unicyb.kiev.ua

Определение позиций изменения диктора в речевом сигнале

В статье рассматривается один из подходов к определению позиции изменения диктора в непрерывном голосовом сигнале. Предложенный подход базируется на использовании коэффициентов мэл-кепстр для построения характеристического вектора и решении о наличии или отсутствии точки изменения диктора на основе меры различия множеств характеристических векторов.

Введение

Во многих задачах, связанных с обработкой речевых сигналов и распознаванием речевой информации, необходимо знать, в каких местах речевого сигнала происходит изменение диктора – лица, чей голос звучит в определенный промежуток времени. В частности в задачах автоматизированного стенографирования [1-3] полезно использовать информацию о смене диктора для более качественной сегментации сигнала. В задачах дикторонезависимого распознавания речи информация о смене диктора позволяет системе адаптироваться под нового диктора.

Задача определения точки изменения диктора состоит в том, чтобы определить позиции в звуковом сигнале, где происходит изменение диктора без информации о дикторах, известной априори. Отсутствие предварительной информации о дикторе отличает задачу определения изменения диктора от более традиционных задач распознавания или верификации диктора. Если бы предварительная информация о дикторе была доступной, для решения задачи можно было бы применить традиционные методы идентификации и распознавания, такие, как методы линейной и нелинейной классификации и методы искусственных нейронных сетей. В реальных же задачах сегментации звукового сигнала получить предварительную информацию о дикторах для составления обучающей выборки и даже информацию о количестве различных дикторов в звуковом сигнале не представляется возможным.

Важным аспектом задачи определения смены диктора является возможность решения задачи в реальном или квазиреальном времени, следовательно вычислительная сложность алгоритма определения смены диктора должна быть относительно невысокой, чтобы решение задачи в реальном времени было возможным на широкораспространенном аппаратном обеспечении.

Существует набор методов определения смены диктора в речевом сигнале [4-6], которые в большинстве своем базируются на использовании коэффициентов мэл-кепстр для построения характеристических векторов, но при этом используют разные подходы для определения степени различия между множествами характеристических векторов или между отдельными характеристическими векторами. В частности в [6] в качестве меры различия предложена взвешенная мера, основанная на взвешенном Евклидовом расстоянии между векторами. При таком подходе для определения

изменения диктора проводится сравнение двух соседних сегментов, причем каждый из сегментов представлен одним характеристическим вектором, который является усредненным характеристическим вектором для сегмента, помноженным на весовой коэффициент, зависящий от класса, к которому принадлежит рассматриваемый вектор. Недостаток такого подхода состоит в том, что случайные возмущения сигнала в сегменте могут существенно исказить усредненный вектор, во избежание чего необходимо проводить качественную нормализацию сигнала, что в реальных условиях не всегда осуществимо или целесообразно.

Другой метод рассматривается в [4], [5]. Для определения точки изменения диктора авторы предлагают использовать меру дивергенции для определения расстояния между отдельно взятыми характеристическими векторами. Несмотря на то, что такой метод дает достаточно высокую точность и является не очень требовательным к вычислительным ресурсам, ввиду того, что в каждый момент времени рассматриваются лишь несколько соседних характеристических векторов, вероятность ошибочного определения точки изменения диктора достаточно высока из-за возможных случайных возмущений в сигнале, локального изменения интонации и т.п. Авторы [4] предлагают дополнить характеристический вектор кроме коэффициентов мэл-кепстр, еще и коэффициентами линейного предсказания и питчем. Несмотря на то, что такой подход дает более широкое представление о голосовом сигнале в характеристическом векторе, он требует значительно большего числа вычислений для расчета дополнительных коэффициентов, что усложняет решение задачи в реальном времени.

В данной статье предлагается еще один подход для определения точки изменения диктора в реальном времени. Предполагается, что на вход системы подается звуковой сигнал, содержащий голосовую информацию, прошедший предварительную обработку для снижения уровня посторонних шумов. Количество дикторов в сигнале, число точек изменения диктора заранее неизвестны. Любая информация о характеристиках дикторов априори также неизвестна. Рассматривается выбор и построение характеристического вектора, приводится мера различия между множествами характеристических векторов и решение о присутствии изменения диктора в заданной точке на основании такой меры. Обсуждаются вопросы применения предложенного алгоритма в системе автоматизированного стенографирования.

Выбор и построение характеристического вектора

Выбор характеристик для задачи определения изменения диктора аналогичен выбору характеристик для задачи идентификации и верификации диктора. Исследования показали, что для задач распознавания диктора одной из самых подходящих характеристик являются коэффициенты мэл-кепстр [7].

Коэффициенты мэл-кепстр определяются как кепстр в области действительных чисел кратковременного звукового сигнала, полученный из преобразования Фурье этого сигнала. Отличие от простого кепстра состоит в том, что для разложения используется нелинейная шкала частот, которая приблизительно описывает особенности слухового восприятия информации человеком.

Полагая, что дискретное преобразование Фурье входного сигнала задано

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk}, \, 0 \le k < N, \qquad (1)$$

определяется набор M фильтров ($m = 1, 2, ..., M$), где фильтр m – это треугольный

фильтр, заданный как:

$$H_m[k]=\begin{cases}0, k<f[m-1]\\ \dfrac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, f[m-1]\le k\le f[m]\\ \\ \dfrac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])}, f[m]\le k\le f[m+1]\\ 0, k>f[m+1]\end{cases}. \qquad (2)$$

Такие фильтры вычисляют средний спектр вокруг каждой из центральных частот с возрастающей шириной, как показано на рис. 1:

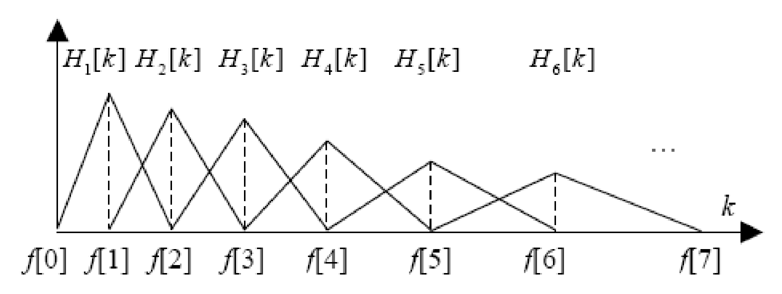


Рисунок 1 – Набор треугольных фильтров для вычисления мэл-кепстра

Пусть f_l и f_h – соответственно самая низкая и самая высокая частоты в наборе фильтров, заданные в Гц, F_s – частота дискретизации в Гц, M – количество фильтров в наборе, N – размер БПФ. Граничные точки фильтров $f[m]$ тогда равномерно расположены по мэл-шкале:

$$f[m]=\left(\frac{N}{F_s}\right)B^{-1}\left(B(f_l)+m\frac{B(f_h)-B(f_l)}{M+1}\right), \qquad (3)$$

где

$$B^{-1}(b)=700\left(\exp\left(\frac{b}{1125}-1\right)\right). \qquad (4)$$

Как правило, для задач анализа голосовых сигналов используется M в пределах от 24 до 40, при этом при расчетах учитываются первые 13 коэффициентов мэл-кепстр [8].

При экспериментальной реализации системы алгоритм построения характеристических векторов был реализован следующим образом: для вычисления мэл-кепстр проходим по сигналу окном типа Хэннинга длиной 1024 сэмплов (0,023 с при частоте дискретизации 44 100 Гц). Начало каждого следующего окна смещено на 10 мс от начала предыдущего. Так, для участка звукового сигнала, где происходит изменение диктора, коэффициенты мэл-кепстр на графике выглядят следующим образом:

На рис. 2 представлен график изменения коэффициентов мэл-кепстр со временем в звуковом сигнале. Прямоугольником выделена область, где происходит смена диктора. На графике можно достаточно отчетливо увидеть различие между коэффициентами мэл-кепстр в левой (до точки смены диктора) и правой (после точки смены диктора) части графика.

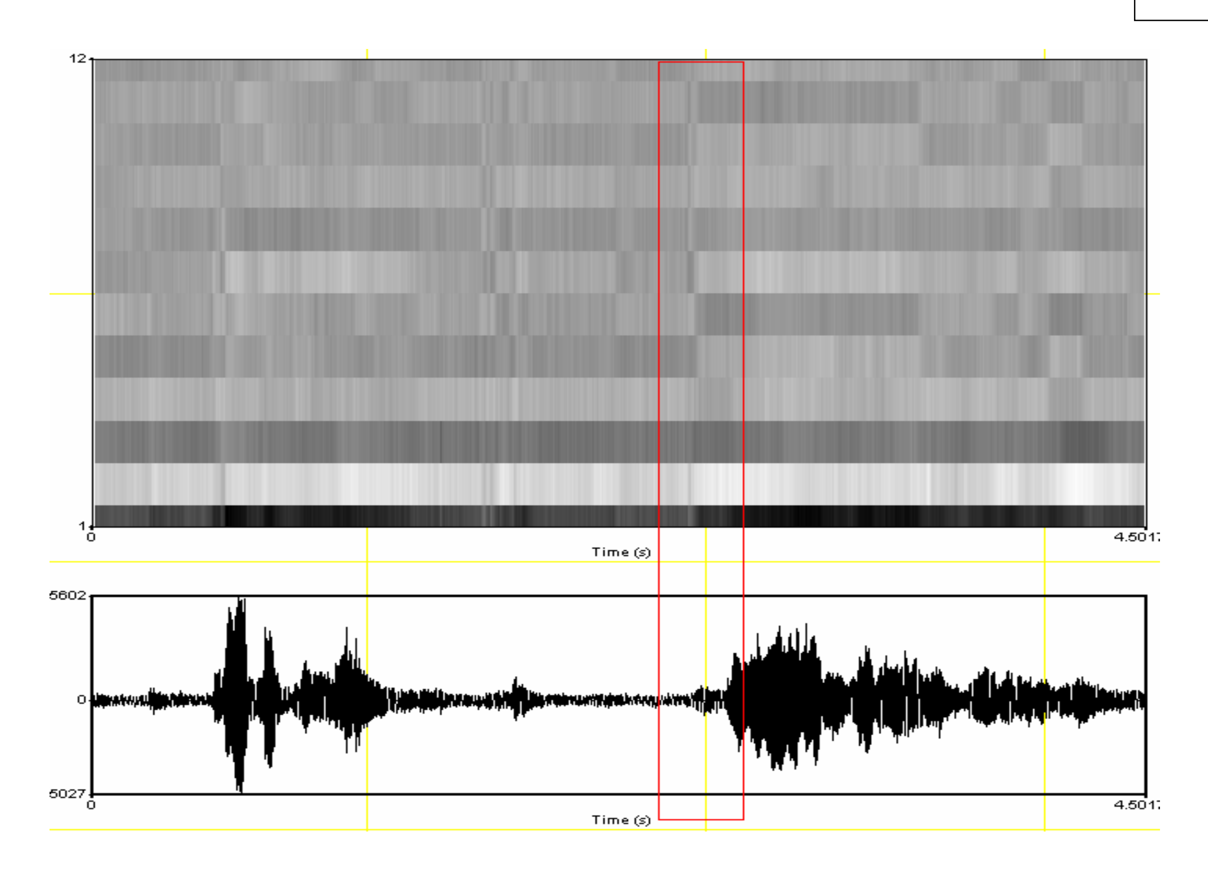


Рисунок 2 – Изменение коэффициентов мэл-кепстр при смене диктора

Мера различия между множествами характеристических векторов

В реальном голосовом сигнале изменение диктора с достаточно большой вероятностью происходит в окружении одной из областей, где в сигнале присутствует пауза. Таким образом для нахождения точек изменения диктора в голосовом сигнале достаточно найти все паузы, в окружении которых возможно изменение диктора, построить множества характеристических векторов сигнала до и после паузы и сравнить эти множества между собой для принятия решения о том, есть ли в окружении данной паузы изменение диктора. Решение о наличии или отсутствии смены диктора принимается на основе вычисления различия между собой множеств характеристических векторов до и после паузы. Если условное расстояние между множествами превышает вычисленный экспериментальным путем порог, то в окружении данной паузы вероятнее всего есть смена диктора.

Паузы в сигнале находятся аналогичным образом, как и в задаче сегментирования звукового сигнала для системы распределенного автоматизированного стенографирования [1]: по сигналу проходим прямоугольным окном заданной длины и вычисляем дисперсию амплитуды сигнала в рамках данного окна. Если дисперсия в рамках данного окна не превышает вычисленного экспериментальным путем порогового значения, то утверждаем, что в соответствующей области сигнала есть пауза.

$$D(x_{i,}x_{i+1},\ldots,x_{i+k}) < \delta_1 . \quad (5)$$

Здесь i – начало окна, для которого проводится анализ, k – длина окна, а δ_1 – экспериментально определенное пороговое значение.

Несколько следующих подряд областей сигнала, в которых дисперсия не превышает порогового значения, объединяются в одну паузу.

Пусть X_1 – множество характеристических векторов звукового сигнала до паузы, а X_2 – множество характеристических векторов сигнала после паузы. Тогда предположение о том, что в окружении данной паузы есть смена диктора возможно, исходя из того, что

$$d(X_1, X_2) > \delta_2, \quad (6)$$

где $d(X_1, X_2)$ – мера различия между множествами векторов, а δ_2 – экспериментально определенный порог.

Меру различия между множествами определим как медиану расстояний между всеми векторами каждого из сравниваемых:

$$d(X_1, X_2) = \mu_{1/2}(d(x_{1i}, x_{2j})) \qquad \forall x_{1i} \in X_1, x_{2j} \in X_2. \quad (7)$$

В качестве расстояния между векторами можно использовать обыкновенное Евклидово расстояние между векторами:

$$d(x_{1i}, x_{2j}) = \sqrt{\sum_{n=1}^{N}(x_{1i}[n] - x_{2j}[n])^2}. \quad (8)$$

Такая мера позволяет определить, насколько разные компоненты содержатся в каждом из множеств, определив, насколько далеко друг от друга находятся векторы каждого из множеств при помощи Евклидового расстояния. Использование медианы в качестве усредненного расстояния позволяет исключить возможные возмущения звукового сигнала в одном из множеств, которые бы могли дать слишком большое или слишком маленькое расстояния между одним или несколькими векторами из одного множества и векторами другого множества. Таким образом значительно уменьшается необходимость в предварительной нормализации звукового сигнала и избавлении его от случайных возмущений, которые могут создаваться звукозаписывающей аппаратурой или случайными посторонними шумами.

Пороговое значение подбирается вручную в результате экспериментов таким образом, чтобы уменьшить количество неверно определенных точек изменения диктора и увеличить количество правильно определенных точек. Как правило, для вычисления порога достаточно проанализировать небольшой участок сигнала, где есть изменение диктора, в дальнейшем порог можно уточнять по мере появления новых дикторов.

Реализация и экспериментальная проверка

Описанный выше метод был реализован в рамках системы автоматизированного стенографирования для сегментации звукового сигнала по точкам изменения диктора. Вычисление точек изменения диктора происходит одновременно с нахождением пауз в звуковом сигнале и сегментации сигнала по паузам. Значение порогового параметра устанавливается вручную при конфигурации системы.

Для проверки эффективности метода было проведено несколько испытаний на реальных звуковых сигналах англоязычным и украиноязычным текстом (для англоязычных текстов использовались фрагменты обучающих фильмов, для украиноязычных – записи фонограмм заседаний ученых советов по защите диссертаций Инсти-

тута кибернетики НАН Украины им. Глушкова). В результате экспериментов выяснились следующие особенности рассматриваемого подхода:

1. Предложенный метод в целом дает достаточно точное распознавание точек смены диктора в различных условиях. Количество пропущенных точек смены диктора как правило не превышало 10 – 15% при правильном подборе пороговых значений.

2. Подбор пороговых значений является нетривиальной задачей и требует достаточно точного определения порога вручную при конфигурации системы, при этом порог часто требует корректировки для различных участков сигнала.

3. Несмотря на то, что точность определения смены диктора достаточно велика, при определенных особенностях звукового сигнала количество неверно определенных точек смены диктора (когда система указывала, что в данной точке есть смена диктора, когда ее там на самом деле нет) может быть также достаточно большим. К таким особенностям звукового сигнала следует отнести существенное изменение интонации одним и тем же диктором, существенное изменение амплитуды сигнала и т.п.

4. Предложенный алгоритм достаточно чувствителен к точному определению пауз в сигнале. Если в голосовом сигнале присутствует фоновая музыка или сильный фоновый шум, правильно определить паузы достаточно сложно, что в свою очередь отрицательно сказывается на количестве правильно определенных точек изменения диктора.

5. Слишком длинные паузы, которые в середине могут содержать возмущение звукового сигнала (шум), также отрицательно сказываются на качестве определения точек смены диктора, так как случайные шумы могут быть восприняты алгоритмом как участок сигнала, содержащий голосовую информацию.

Выводы и дальнейшее развитие предложенного подхода

Несмотря на вышеперечисленные некоторые недостатки рассматриваемого подхода, точность определения точек смены диктора достаточна для большинства применений, включая задачу сегментации сигнала в системе автоматизированного стенографирования. Качество определения точек изменения диктора при предложенном подходе можно повысить за счет следующего:

1. Качественной подготовки звукового сигнала перед его сегментацией, в частности, избавлением сигнала от посторонних шумов.

2. Автоматизированного определения пороговых значений для различных участков звукового сигнала.

3. Расширения характеристических векторов за счет добавления к ним, например, информации о частоте основного тона сигнала (питча), что может повысить качество определения точек изменения диктора, особенно, когда происходит смена мужского голоса на женский и наоборот.

Среди преимуществ рассматриваемого подхода также следует отметить его относительно невысокую требовательность к вычислительным ресурсам, что позволяет применять его для решения задачи в реальном времени.

Литература

1. Информационная система распределенного компьютерного документирования речевых фонограмм заседаний / Ю.Г. Кривонос, Ю.В. Крак, А.В. Бармак, А.С. Загвздин // Управляющие системы и машины. – 2008. – № 3.

2. Розподілене комп’ютерне документування голосових мовних фонограм / Ю.Г. Кривонос, Ю.В. Крак, О.В. Бармак, О.С. Загваздін // Проблеми програмування. – 2008. – № 2 – 3.
3. Автоматизированная система стенографирования / Ю.Г. Кривонос, Ю.В. Крак, А.В. Бармак, А.С. Загваздин // Искусственный интеллект. – 2009. – № 3. – С. 228-233.
4. Lu L. Speaker change detection and tracking in real-time news broadcasting analysis [Электронный ресурс] / L. Lu, H.-J. Zhang // Proceedings of the tenth ACM international conference on Multimedia. – December 1 – 6, 2002. – Juan les Pins, France ACM, 2002. – Режим доступа : www.informatik.uni-trier.de/~ley/db/conf/mm/index.html
5. Universal background models for real-time speaker change detection [Электронный ресурс] / T.Y. Wu, L. Lu, K. Chen, H.-J. Zhang // Microsoft Research. – Режим доступа : http://research.microsoft.com/users/llu/publications/mmm03_ubmforspkseg.pdf.
6. Kwon. Speaker change detection using a new weighted distance measure / Kwon, Narayanan // International conference on spoken language processing. – 2002. – Vol. 4. – P. 2078-2086.
7. Reynolds D.A. Robust text-independent speaker identification using Gaussian mixture speaker models / D.A. Reynolds, R.C. Rose // IEEE transactions on speech and audio processing. – 1995. – Vol. 3, № 1. – P. 238-246.
8. Huang X. Spoken language processing: a guide to theory, algorithm and system development / X. Huang, A. Acero, H.W. – HonPrentice Hall, 2001.
9. Ajmera J. Robust speaker change detection / J. Ajmera, I. McCowan, H. Bourlard // IEEE signal processing letters. – 2004. – Vol. 11, № 8. – P. 689-695.
10. Saha G. Modified Mel-frequency cepstral coefficient [Электронный ресурс] / G. Saha, U.S. Yadhunandan // Department of electronics and electrical engineering. – Technical university of Denmark, 2004. – Режим доступа : recherche.ircam.fr/equipes/analyse-synthese/burred/phd/burred_phd.pdf.

Ю.Г. Кривонос, Ю.В. Крак, О.С. Загваздін

Визначення зміни диктора у мовному сигналі
У статті розглядається один з підходів до визначення позиції зміни диктора у неперервному мовному сигналі. Запропонований підхід базується на використанні коефіцієнтів мел-кепстр для побудови характеристичного вектора і прийнятті рішення про існування чи відсутність точки зміни диктора на основі запропонованої міри відмінності множин характеристичних векторів.

Yu.G. Kryvonos, Yu.V. Krak, O.S. Zagvazdin

Detect Speaker Change in Continuous Speech Signal
One of the approaches to detect speaker change in continuous speech signal is proposed in the paper. Suggested approach is based on using the mel-frequency cepstral coefficients to build a characteristic vector. Decision on existence or absence of speaker change at a given point is based on a proposed dissimilarity measure between the sets of characteristic vectors

Статья поступила в редакцию 21.06.2010.